Statistiques appliquées (L3)
Groupes de TDs 14 et 15

Correction détaillée de l'interrogation écrite n°1

Exercice 1:

1- On veut vérifier que $f_{X,Y}$ est une fonction de densité de probabilité sur R^2. Il s'agit (essentiellement) de démontrer que:
a- $f_{X,Y}(x,y) \geq 0$ pour tout $(x,y) \in R^2$: cette propriété est évidemment vérifiée.
b- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$: cette propriété est aisément vérifiable du moment qu'on ne se trompe pas dans le calcul de l'intégrale. Il suffit d'écrire que:

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy &= \int_0^{\sqrt{2}} \int_0^1 f_{X,Y}(x,y)dxdy \quad \text{car } f_{X,Y} \text{ est nulle en dehors de } [0,1] \times \left[0, \sqrt{2}\right] \\ &= \int_0^{\sqrt{2}} \int_0^1 3x^2 y dxdy \\ &= 3 \int_0^{\sqrt{2}} \left(\int_0^1 x^2 dx \right) ydy \end{aligned}$$

or, d'une part, $\int_0^1 x^2 dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}$ donc

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = \int_0^{\sqrt{2}} ydy$$

et d'autre part, $\int_0^{\sqrt{2}} ydy = \left[\frac{y^2}{2}\right]_0^{\sqrt{2}} = 1$ donc

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$$

2- La densité marginale de X au point $x \in R$ est définie par:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy$$

Lorsque $x \notin [0,1]$ on sait que $f_{X,Y}(x,y) = 0$ pour tout $y \in R$, ce qui permet d'écrire que $f_X(x) = 0$.
Lorsque $x \in [0,1]$ $f_{X,Y}(x,y) = \begin{cases} 3x^2y & \text{si} \quad y \in \left[0, \sqrt{2}\right] \\ 0 & \text{sinon} \end{cases}$ d'où

$$f_X(x) = \int_0^{\sqrt{2}} 3x^2 ydy = 3x^2 \int_0^{\sqrt{2}} ydy = 3x^2$$

Ainsi,

$$f_X(x) = \begin{cases} 3x^2 & \text{si} \quad x \in [0,1] \\ 0 & \text{sinon} \end{cases}$$

La densité marginale de Y au point $y \in R$ est définie par:

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx$$

Un raisonnement similaire à celui suivi pour le calcul de f_X permet d'aboutir à :

$$f_Y(y) = \begin{cases} y & \text{si} \quad y \in \left[0, \sqrt{2}\right] \\ 0 & \text{sinon} \end{cases}$$

3- Il est aisé de voir que pour tout $(x, y) \in R^2$:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

ce qui permet d'affirmer que les variables aléatoires X et Y sont indépendantes.

Exercice 2

1- Rappel: si la variable aléatoire X suit une loi uniforme sur l'intervalle $[a, b]$ alors sa fonction de densité est définie par:

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & \text{si} \quad x \in [a, b] \\ 0 & \text{sinon} \end{cases}$$

Il en découle que la fonction densité de X_i est définie par

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & \text{si} \quad x \in [\theta, 2\theta] \\ 0 & \text{sinon} \end{cases}$$

Le calcul de l'espérance et de la variance de X_i peut se faire de deux manières:

- ou bien on connaît l'espérance et la variance (classiques!) d'une variable aléatoire X tirée dans une loi uniforme sur un intervalle $[a, b]$, à savoir $E(X) = \frac{a+b}{2}$ et $V(X) = \frac{(b-a)^{12}}{2}$ et on applique tout simplement la "formule". On trouve alors $m = \frac{3\theta}{2}$ et $\sigma^2 = \frac{\theta^2}{12}$

- ou bien on ne connaît pas la "formule" en question et on utilise alors la définition de l'espérance et de la variance d'une variable aléatoire continue. Par définition de l'espérance, on a :

$$\begin{aligned} m &= \int_{-\infty}^{+\infty} x f_\theta(x) dx \\ &= \int_\theta^{2\theta} x \frac{1}{\theta} dx \\ &= \frac{1}{\theta} \left[\frac{x^2}{2} \right]_\theta^{2\theta} \\ &= \frac{1}{\theta} \frac{3\theta^2}{2} \\ &= \frac{3\theta}{2} \end{aligned}$$

Et par définition de la variance, on a :

$$\begin{aligned} \sigma^2 &= E(X_i^2) - E(X_i)^2 \\ &= \int_{-\infty}^{+\infty} x^2 f_\theta(x) dx - m^2 \\ &= \int_\theta^{2\theta} x^2 \frac{1}{\theta} dx - m^2 \\ &= \frac{1}{\theta} \left[\frac{x^3}{3} \right]_\theta^{2\theta} - \frac{9\theta^2}{4} \\ &= \frac{1}{\theta} \frac{7\theta^3}{3} - \frac{9\theta^2}{4} \\ &= (\frac{7}{3} - \frac{9}{4}) \theta^2 \\ &= \frac{\theta^2}{12} \end{aligned}$$

2- La moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur convergent de l'espérance m (d'après la loi des grands nombres). Il s'agit d'un estimateur sans biais (le fait que $E(\bar{X}_n) = m$ découle immédiatement du caractère linéaire de l'espérance).

Remarque: Dans le cas d'une loi uniforme sur un intervalle, la moyenne empirique n'est pas le "meilleur" estimateur de l'espérance (on peut établir qu'il ne s'agit pas d'un estimateur de variance minimale...mais ceci n'est pas trivial).

Afin de déduire un estimateur convergent de θ, il suffit de remarquer que $\theta = \frac{2}{3}m$ (d'après la question précédente). On peut alors affirmer que $\hat{\theta}_n = \frac{2}{3}\bar{X}_n$ est un estimateur convergent du paramètre θ.

3- $E(\hat{\theta}_n) = E(\frac{2}{3}\bar{X}_n) = \frac{2}{3}E(\bar{X}_n) = \frac{2}{3}m$

$V(\hat{\theta}_n) = V(\frac{2}{3}\bar{X}_n) = \frac{4}{9}V(\bar{X}_n)$

or $V(\bar{X}_n) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} = \frac{\theta^2}{12n}$ (la deuxième égalité découle du fait que les X_i sont indépendants; si ça n'avait pas été le cas, il aurait fallu tenir compte des covariances)

donc

$$V(\hat{\theta}_n) = \frac{4}{9}\frac{\theta^2}{12n} = \frac{\theta^2}{27n}$$

4- D'après le TCL:

$$\sqrt{n}(\bar{X}_n - m) \xrightarrow{loi} N(0, \sigma^2)$$

donc, pour n suffisamment grand (on suppose que c'est le cas ici) on peut considérer que $\bar{X}_n$ suit approximativement la loi $N(m, \frac{\sigma^2}{n})$, c'est-à-dire $N(\frac{3\theta}{2}, \frac{\theta^2}{12n})$. Ceci implique que $\hat{\theta}_n (= \frac{2}{3}\bar{X}_n\ )$ suit approximativement la loi normale $N(\frac{2}{3}\frac{3\theta}{2}, \frac{4}{9}\frac{\theta^2}{12n})$ c'est-à-dire $N(\theta, \frac{\theta^2}{27n})$.

On retrouve bien que les paramètres de la loi normale approchant la loi de $\hat{\theta}_n$ sont l'espérance et la variance de $\hat{\theta}_n$ précédemment calculées.

Pour déterminer un intervalle de confiance bilatéral à 95% pour θ, il suffit d'utiliser l'approximation normale que nous venons de faire. En effet, elle peut être ré-écrite sous la forme:

$$U_n = \frac{\hat{\theta}_n - \theta}{\sqrt{\frac{\theta^2}{27n}}} = \frac{3\sqrt{3n}\left(\hat{\theta}_n - \theta\right)}{\theta} \backsim N(0,1)$$

Ceci implique que :

$$P(-1,96 \leq \frac{3\sqrt{3n}\left(\hat{\theta}_n - \theta\right)}{\theta} \leq 1,96) = 0.95$$

qu'on peut ré-écrire sous la forme

$$P(-1,96 \leq 3\sqrt{3n}\left(\frac{\hat{\theta}_n}{\theta} - 1\right) \leq 1,96) = 0.95$$

ou encore

$$P(1 - \frac{1,96}{3\sqrt{3n}} \leq \frac{\hat{\theta}_n}{\theta} \leq 1 + \frac{1,96}{3\sqrt{3n}}) = 0.95$$

c'est-à-dire

$$P(\frac{1}{1 + \frac{1,96}{3\sqrt{3n}}} \leq \frac{\theta}{\hat{\theta}_n} \leq \frac{1}{1 - \frac{1,96}{3\sqrt{3n}}}) = 0.95$$

d'où

$$P(\frac{\hat{\theta}_n}{1 + \frac{1,96}{3\sqrt{3n}}} \leq \theta \leq \frac{\hat{\theta}_n}{1 - \frac{1,96}{3\sqrt{3n}}}) = 0.95$$

A.N: Pour $n = 500$ et une valeur observée de 10 pour l'estimateur $\bar{X}_n$ (qui implique une valeur observée de $\frac{20}{3}$ pour $\hat{\theta}_n$) on trouve l'intervalle $[6,556\ ,\ 6,781]$.

Exercice 3

1- Les variables aléatoires X_i et X_{i+1} peuvent prendre la valeur 0 ou la valeur 1 donc leur produit $Y_i = X_i X_{i+1}$ peut prendre la valeur 0 (c'est le cas si au moins l'une des variables X_i et X_{i+1} prend la valeur 0) ou la valeur 1 (c'est le cas lorsque les variables X_i et X_{i+1} prennent toutes les deux la valeur 1).

2- Vu que $X_i \in \{0,1\}$ et $X_{i+1} \in \{0,1\}$, il est clair que $Y_i = 1 \iff X_i = 1$ et $X_{i+1} = 1$, d'où $P(Y_i = 1) = P(X_i = 1, X_{i+1} = 1)$. Or les variables X_i et X_{i+1} sont indépendantes donc $P(Y_i = 1) = P(X_i = 1) P(X_{i+1} = 1)$. X_i et X_{i+1} étant des variables de Bernouilli de paramètre p, il s'ensuit que pour tout $i \in N$:

$$P(Y_i = 1) = p^2$$

Y_i prend ses valeurs dans l'ensemble $\{0,1\}$ et $P(Y_i = 1) = p^2$ donc Y_i suit une loi de Bernouilli de paramètre p^2. Son espérance est donc p^2 et sa variance est $p^2\left(1-p^2\right)$.

3-a- Par linéarité de l'espérance,

$$E(S_n) = \frac{\sum\limits_{i=1}^{n} E(Y_i)}{n} = \frac{np^2}{n} = p^2$$

3-b- On a vu que $Y_i = 1 \iff X_i = 1$ et $X_{i+1} = 1$. De même, on a : $Y_{i+1} = 1 \iff X_{i+1} = 1$ et $X_{i+2} = 1$. Il s'ensuit que

$$Y_i = 1 \text{ et } Y_{i+1} = 1 \iff X_i = 1\ ,\ X_{i+1} = 1 \text{ et } X_{i+2} = 1$$

donc

$$P(Y_i = 1, Y_{i+1} = 1) = P(X_i = 1, X_{i+1} = 1, X_{i+2} = 1)$$

or les variables X_i, X_{i+1} et X_{i+2} sont indépendantes donc

$$P(Y_i = 1, Y_{i+1} = 1) = P(X_i = 1) P(X_{i+1} = 1) P(X_{i+2} = 1)$$

c'est-à-dire que

$$P(Y_i = 1, Y_{i+1} = 1) = p^3$$

On remarque que $P(Y_i = 1, Y_{i+1} = 1) \neq P(Y_i = 1)P(Y_{i+1} = 1)$. En effet, le premier terme est égal à p^3 alors que le second terme est égal à $p^2 \times p^2 = p^4$ (on utilise le résultat de la question 2). On peut alors affirmer que les variables Y_i et Y_{i+1} ne sont pas indépendantes.

3-c- Si les variables $Y_1, Y_2, ..., Y_n$ étaient indépendantes alors on pourrait écrire *directement* l'égalité

$$Var(S_n) = \frac{1}{n^2}\left(Var(Y_1) + Var(Y_2) + ... + Var(Y_n)\right)$$

car les termes de covariance auraient été nuls. Or d'après la question précédente, ces variables ne sont pas indépendantes puisque, pout tout $i \in N$, les variables Y_i et Y_{i+1} ne sont pas indépendantes. Il s'ensuit qu'on ne peut pas écrire *directement* l'égalité en question: il faut faire le calcul en tenant compte des termes de covariance.

3-d- D'après l'inégalité de Tchebychev, on a:

$$P(\left|S_n - E(S_{n)}\right| > \varepsilon) \leq \frac{Var(S_n)}{\varepsilon^2}$$

or $E(S_n) = p^2$ et $Var(S_n) = \frac{p^2(1-p^2)}{n} + \frac{2(n-1)}{n^2}(p^3 - p^4) \to 0$ lorsque $n \to +\infty$ donc

$$\lim_{n \to +\infty} P(\left|S_n - p^2\right| > \varepsilon) = 0$$

On vient de redémontrer dans un cas particulier la loi (faible) des grands nombres.